Audiovisual Integration Under Different Conditions of Hearing Loss

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation *with honors distinction* in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

By

Devon Milkie

The Ohio State University

April 2014

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

In any listening environment, normal or compromised, humans integrate the auditory and visual cues provided, in comprehending speech. One unresolved question is how different forms of hearing loss differentially impact the integration process. The present study investigated how degradation of the auditory signal due to two types of hearing loss inhibited a listener's ability to integrate. Ten adult listeners, with normal or corrected-to-normal vision and auditory thresholds at or better than 25 dB HL across all frequencies, were presented with everyday sentences produced by four different talkers from the HeLPs software by Sensimetrics, Inc. Each sentence was presented in audio-only, visual-only, and audio+visual modalities. Auditory input simulated a sloping hearing loss (55 dB HL at 1000 Hz) and the stimulus presented by an 8-channel cochlear implant. Results of testing suggest that sentences presented in the cochlear implant condition were more intelligible, while sentences in the sloping condition showed the greatest audio-visual integration. These findings raise a question about the fidelity of the cochlear implant simulation in the software, given that such a result is not likely in real-world situations. Results of the present study may have implications for development of speech-reading and aural rehabilitation programs in the future.

Acknowledgements

I would like to thank my advisor, Dr. Janet M. Weisenberger, for her guidance through the process of fulfilling my honors thesis requirements. Her patience and passion for research have inspired me to learn more both in and outside of my discipline. Additionally, I would like to thank my participants for their time and willingness to participate.

This project was supported by an SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

For individuals with normal hearing, speech perception is considered to involve a predominantly auditory signal. But when that signal is undermined, in the case of a hearing impaired individual or in a noisy environment, both the auditory and visual modalities can be employed to better interpret the signal. Visual cues from the talker provide information to fill the void, supporting speech perception as a multi-modal process. A study by McGurk and MacDonald (1976) asserts that audio-visual integration occurs even when the acoustic speech signal is perfectly intelligible. This concept, known as the McGurk effect, prevails even after subjects are made aware of the effect. Subjects in the McGurk and MacDonald (1976) study exhibited either a fused or a combination response. A "fused" response occurs when the information from the audio and visual modalities transforms to produce a speech sound not presented in either modality. For example, the auditory signal /pa/ is dubbed over the visual signal /ka/. A significant number of subjects reported hearing the sound /ta/. Conversely, a "combination" response is one where "relatively unmodified" information from the audio and visual modalities are heard in succession. For example, the auditory signal /ga/ is dubbed over the visual signal /ba/. Subjects reported perceiving /gabga/, /bagba/, or other arrangements of the sounds presented. The results of this study suggest that audio-visual integration occurs consistently throughout speech perception.

Since the introduction of the McGurk effect, research has sought to determine the information in the visual signal that is provided to a listener. A study by Jackson (1988) observed that although the visual signal reliably provides only place of articulation information to the recipient, the place of articulation, rate of articulation, and oral cavity

shape provide tangible markers with which humans speech read. A listener's ability to improve the intelligibility of an auditory signal with a visual signal may be affected by the effectiveness of a talker's articulators. Jackson (1988) also studied the importance of aspects of the visual signal in the context of both isolated utterances and coarticulation, the change in production of one sound due to the production of an adjacent sound. She determined that visual cues, such as lip extension, rounding, and separation, are crucial to discerning between different phonemes while speech reading, and that visual coarticulation effects can facilitate or impair the speech-reading process. Grant and Braida (1991) determined that speech intelligibility in individuals with hearing loss can be improved by including a visual signal with their compromised auditory signal.

A study by Munhall et al. (2004) reiterated the importance of the visual signal found by Jackson (1988), due to the context information that may be gathered from "the talker's identity, emotional and physical state, focus and attention, and degree of social and conversational engagement, as well as information about the utterance" (Munhall et al. 2004, p. 575). Additionally, Munhall et al. (2004) suggested that the visual signal could still yield most necessary cues even when the visual signal is significantly degraded, given that it is accompanied by an auditory signal. They altered the spatial frequencies of their talker video in several ways, which led to the conclusion that most information from the visual signal is gathered from the low to mid-frequency portion of the image. Even with a normal visual signal, Munhall et al. (2004) stated that human observers with normal hearing do not use all information available from facial images. Articulation cues provided by the structures of the face and mouth may be useful as temporal cues during cross-modal integration. MacDonald et al. (2000) agree that a visual

signal may be more useful for temporal cues by observing the prevalence of the McGurk effect after spatially degrading the video image. An illusory response was recorded even at the highest level of degradation. The results of Munhall et al. (2004) and MacDonald et al. (2000) suggest that some of the specifics of a visual signal may not be as necessary as previously thought, but their findings may have interesting implications for integration opportunities where the auditory stimulus is already degraded. The visual signal could provide the temporal cues necessary to make the auditory signal more intelligible. Conversely, a degraded auditory signal may require the listener to rely more heavily on the visual signal for place and rate of articulation.

Like visual signals, auditory signals also have characteristics for optimal perception. Shannon et al. (1998) built on extensive research in measuring the frequency regions of the speech signal that contribute the most critical information to the speech percept. Their work demonstrated that speech can be perceived under some circumstances with purely temporal cues. Shannon et al. (1998) observed that consonant recognition required no spectral information. They note that consonantal voicing and manner cues are correctly perceived even when spectral cues are reduced to two bands of modulated noise. Vowel recognition, however, depends on spectral information and therefore, is more sensitive to distortion. This study has major practical implications; for example, electrode arrays in cochlear implants can be mapped for optimal neural responses to match these spectral emphases.

In situations where a speech signal is degraded, a speaking style known as “clear speech” often increases speech intelligibility. Clear speech differs from conversational speech in both its linguistic and acoustic properties. Krause and Braida (2002)

characterize clear speech using "a slower speaking rate, greater temporal modulation, increased range of voice fundamental frequency, an expanded vowel space, and more stimulus energy in high frequencies." Several other studies have confirmed these characteristics (Chen, 1980; Picheny et al.,1985; Uchanski et al., 1992; Payton et al.,1994). These studies have shown that individuals whose speech characteristics approximate those of clear speech are more intelligible. Clear speech increases the redundancy of the speech signal, which provides more opportunities for the listener to receive information from it. The increased range of voice fundamental frequency makes clear speech particularly useful in situations where the listener has decreased frequency resolution, like in a cochlear implant patient.

Hearing in a cochlear implant recipient differs from that of a more typical hearing impaired individual because of the type of signal received by the auditory nerve. The tonotopically-organized basilar membrane, if optimally functioning, is sensitive to a range of frequencies from 20 Hz to 20,000 Hz. The basilar membrane in a cochlear implant recipient, however, needs to be activated by electrode arrays spaced throughout the cochlea, which greatly decreases the available frequency resolution of the ear. Acoustic information processed by a cochlear implant electrically stimulates the auditory system "in a manner that produces highly unnatural patterns of neural activity" (Shannon et al., 1998). The most effective cochlear implants have 22 electrodes, or channels, which account for less than 1% of the hair cells in the cochlea used to receive sound signals. Consequently, cochlear implant users receive very little spectral information from speech and relevant noise. Given the observations from the aforementioned studies, cochlear

implant users may benefit from audio-visual integration because they lack most of the spectral cues necessary for discriminating vowels.

The speech signal and how it is degraded, paired with characteristics of the talker and characteristics of the listener, shape the outcome of a communication setting. Hager (2013) deconstructed talker differences in her study, using a simulated configuration of hearing loss (sloping 55 dB HL at 1000 Hz) for presenting the auditory signal. Hager found large differences across talkers in audio-visual integration and speech intelligibility, but looked at only one hearing loss configuration. An unanswered question is whether the same pattern of performance would be seen with a very different hearing loss configuration, such as might be the case for a cochlear implant recipient, where the information delivered about the acoustic signal is quite different. The present study analyzed the difference between integration with a simulated sloping hearing loss (55 dB HL at 1000 Hz) and a simulated 8-channel cochlear implant. Differences observed across the hearing conditions, each modality, and each talker, may support previous work suggesting that the reduced spectral content in the cochlear implant signal will benefit more from the addition of visual cues.

Based on past studies, hypotheses can be made about the outcome of the present study regarding audio-visual integration under different conditions of hearing loss. The results of Grant and Braida (2001) confirmed that hearing impaired individuals show increased speech recognition when their degraded auditory signal is accompanied by visual signal. The work of Munhall et al. (2004) and MacDonald et al. (2000) supported the hypothesis that the aforementioned visual signal is mostly useful for temporal cues. Results of testing in the present study were expected to suggest that sentences presented

in the sloping condition were more intelligible, while sentences in the cochlear implant condition showed the greatest audio-visual integration.

Chapter 2: Method

Participants

Participants in this study included 10 listeners, 5 male and 5 female, with normal hearing, characterized by thresholds of 25 dB HL or better, across the frequencies of 250-4000 Hz as measured by an audiometric test. Additionally, each participant reported normal or corrected-to-normal vision. All 10 listeners were native speakers of American English. Participants were compensated $10 per hour for dedicating their time and effort to this study.

Stimulus Presentation

HeLPS

The stimuli used in the present study consisted of 368 randomized and prerecorded sentences programmed into the HeLPS software. HeLPS—Hearing Loss and Prosthesis Simulator (Sensimetrics Corporation)— is a computer software program that simulates the auditory communication difficulties associated with hearing loss along with the possible benefits provided by hearing aids and cochlear implants. HeLPS software can simulate hearing loss of any configuration and degree and differentiate between air and bone conduction thresholds for the right and/or left ears, and can also simulate tonal or noisy tinnitus. Sensimetrics Corporation provides a graphic interface on the computer for controlling the simulation and a set of calibrated headphones for listening to the simulator's output. This program provides presentation in three modalities: audio-only, visual-only, and audio+visual. The characteristics of hearing and prosthesis are specific

to the left and right ears, with the ability to select loss and prosthesis setting, talkers, background noises, and reverberation.

Talkers

Within HeLPS, there are 10 talkers total, 5 male and 5 female. These recorded talkers represent a myriad of ethnicities with a range of 13 to 67 sentences each. The sentences are either statements or questions referencing everyday topics such as events, people, and places. The present study chose 4 talkers (2 male and 2 female) from the set used by Hager (2012). They were selected based on Hager's findings that the talkers produced utterances that yielded high degrees of audiovisual integration.

Presentation

Twenty recorded sentences from each of the chosen talkers were presented to listeners in three modalities: auditory-only, visual-only, and audio+visual. These presentation styles are available by selecting the appropriate option in HeLPS before playing the talker's sentence.

Auditory-only Presentation

For auditory-only presentations, sentences were presented with an auditory stimulus, but no visual stimulus. Ten sentences for each talker were presented at 75 dB under a setting used to simulate a sloping high frequency hearing loss (55dB at1000 Hz), an audiogram of which can be seen in Figure 1. Ten different sentences simulated hearing with an 8-channel cochlear implant. All sentences were presented through Sennheiser supra-aural headphones calibrated specifically for the HeLPS program.

Visual-only Presentation

Using a computer monitor, the visual-only stimulus was presented using a 5x4 in. video image of the talker speaking a sentence. No auditory stimulus accompanied the video image.

Audio+Visual Presentation

Audio+visual presentation consisted of both the audio and visual options outlined above.

Headphones Calibration

Headphones used during testing had previously been calibrated using a KEMAR manikin. By using continuous speech shaped noise at 65 dB SPL, with anechoic settings, a steady noise stimulus (with brief occasional pauses) was presented and measured by the equipment.

Test Enclosure

Testing for this study was performed in a lab room located in the basement of Pressey Hall. To ensure a quiet environment, testing was completed on nights and weekends.

Procedure

After providing informed consent under Ohio State University protocol 2012B0049, each subject was seated in a quiet lab room in the basement of Pressey Hall. The experimenter provided them with written and spoken instructions for the procedure. Using a dual headphone jack splitter and laptop, the participant and experimenter put on headphones. The participant was then given a demonstration of the HeLPS software.

Following the demonstration, subjects were presented with a video-only or audio-only stimulus, then shown the opposite modality, based on random assignment. The third

presentation always consisted of the audio+visual modality. After each presentation, participants were instructed to vocalize their perceived content of the sentence. The examiner then recorded verbatim the listener responses on a score sheet. No feedback regarding the accuracy of their responses was given to the listeners. The presentation procedures were applied to a total of 20 sentences from each of the 4 designated talkers.

The order in which the talkers were presented, along with the determination of the first modality and hearing condition, was based on random assignment across subjects. A replacement algorithm in the software randomized the selection of a sentence for each trial.

To avoid fatigue, frequent breaks were encouraged. Each subject was allotted 1.5 hour appointments based on their availability, with only one appointment allowed per day, for a total of 4 sessions per listener. Total time for each listener was approximately 6 hours.

Chapter 3: Results and Discussion

To analyze results, a 3-factor repeated measures analysis of variance (ANOVA) was performed on percent key words correct. The factors were hearing condition, presentation condition (modality), and talker. Results of the analysis are reported below.

Results indicated a significant main effect for all factors, hearing condition, modality, and talker. Figure 2 shows overall performance for the sloping and CI hearing conditions. There is a significant difference across modalities, $F(2,18)=280.94$, $p<.001$, wherein audio+visual conditions yielded the highest scores and visual-only, the lowest. Across modalities, the CI condition produced slightly better results than the sloping condition, $F(1,9)=8.545$, $p=0.017$.

In addition to hearing condition and modality, data for each of the talkers was analyzed separately using means contrasts, to get a better sense of differences across talkers that contribute to overall findings. Joe proved to be the most intelligible talker, as he was significantly better than Ann (10.967, p=.003), Bob (7.40, p=.006), and Christina (15.283, p=<.001). Bob was more intelligible than Christina (7.883, p=.002). However, Ann was not statistically better than either Bob or Christina.

In addition, the amount of audiovisual integration provided by each talker was calculated as the difference between the audio+visual condition and the best single modality. These results are shown in Table 1. Overall, the male talkers showed far less integration than the female talkers by approximately 10%, as seen in Table 1. Interestingly, these results differ from the findings of Hager (2013), wherein there was no significant difference in integration between the male talkers and female talkers.

Figures 3 and 4 compare the percent key words correct for each talker and each modality. A statistically significant interaction was found between modality and talker, $F(6,54)=12.838$, $p<.001$. Figure 3 shows results for the sloping hearing loss condition, where Joe is the most intelligible talker in all three modalities. Ann is the least intelligible talker for both the audio-only and audio+visual modalities, while Christina and Bob have the lowest intelligibility score in the visual-only modality. In the cochlear implant condition, Joe has the highest intelligibility scores in the visual-only and audio+visual modality. Bob and Joe have the highest score in the audio-only condition. Christina has the lowest scores in all three modalities. Figure 5 collapses the data across modality and hearing condition to analyze the effectiveness of each talker. Overall, Joe is the most intelligible talker and Christina is the least intelligible.

Anecdotal comments gathered from the listeners may further support the significant differences across talker and modality observed in the present study. Several listeners expressed their distaste for Bob and Christina, stating that the large size of their lips impeded the ability to decipher consonant sounds in all modalities. Interestingly, the listeners had very few complaints about Ann, despite the fact that she proved less intelligible than Bob. Joe was the most intelligible speaker, yet the listeners commented that his lack of facial expression made him "boring." Listeners also attempted to connect the content of the sentences with each talker. They often commented that they did not expect the talker to have a child or live in a certain city.

Chapter 4: Summary and Conclusion

Overall, results of the present study indicate that there is a significant difference between conditions of hearing loss, though it did not reflect the hypothesized outcome. There were also significant effects across modality and talker. The data collected supported the findings of McGurk and MacDonald (1976), wherein audio-only stimuli were significantly more intelligible than visual-only stimuli, but audio+visual stimuli were the most intelligible. The results of the present study indicate that male talkers produced less integration than female talkers, although Hager (2013) did not find significant differences based on gender. A significant interaction occurred between hearing condition and modality, in which all modalities were more intelligible in the cochlear implant condition than the sloping hearing loss condition. These results are counterintuitive and may be indicative of a problem in the software algorithm. An additional interaction was found between modality and talker. Joe produced the highest intelligibility scores for all three modalities, while Ann and Christina showed the lowest. There are several explanations for the level of intelligibility across talkers, but further data would be needed to determine the true cause of the above observations.

The HeLPS software presented several challenges that may have skewed the results of the present study. As mentioned above, the data collected would be very unlikely in the real world, leading to questions about the fidelity of the cochlear implant simulation in the software. The HeLPS software provides a different set of sentences for each talker, which prevented us from comparing results across sentence set. The program also randomizes the sentences selected for presentation, so every listener was presented with different sentences within each talker.

The present study gathered results from only 4 talkers and 10 listeners due to time constraints, but future expansion on the project may include more talkers and listeners, as well as additional hearing loss conditions. The present study did not observe the effect of a conductive hearing loss on audiovisual integration, but obtaining that information would give a more well-rounded explanation of why integration differed. In addition to decreasing the intensity of sound, sensorineural hearing loss may distort speech due to broadening of auditory filters. Comparatively, speech processed by a cochlear implant lacks some spectral features, causing it to sound different than speech perceived by normal hearing individuals. A strictly conductive hearing loss would have no distortion aspect, which may produce different results than those found in the present study.

To optimize the efficiency of an aural rehabilitation program, family members of the hearing-impaired patient must be educated on ways to make their speech more intelligible. Although the present study did not address the factors that make speech more intelligible, it does have implications on how hearing condition should weigh on the customization of a patient's rehabilitation program. Our data concluded that individuals with a sloping hearing loss perceive fewer words correctly than those with a cochlear implant. From these conclusions, an aural rehabilitation program for a patient with a sloping hearing loss should put a stronger emphasis on auditory training and speech reading than a program for a patient with a cochlear implant. Although some conclusions were able to be drawn from the data collected, observations about other aspects of integration, like the effectiveness of training, are necessary to improve and personalize aural rehabilitation programs for hearing-impaired patients.

Chapter 5: References

Grant, K.W., and Braida, L.D., (1991). “Evaluating the articulation index for audiovisual input.” Journal of the Acoustical Society of America 89, 2952-2960.

Hager, A. (2013). “Talker differences and gender effects in audio-visual integration.” Undergraduate Thesis, The Ohio State University.

Jackson, P. (1988). “The Theoretical Minimal Unit for Visual Speech Perception: Visemes and Coarticulation.” The Volta Review 90, 99-115.

Krause, Jean C.; Braida, Louis D. (2002). The Journal of the Acoustical Society of America vol. 112 issue. p. 2165-2172

MacDonald, J., Andersen, S., Backmann, T. (2000). “Hearing by eye: how much spatial degradation can be tolerated?” Perception vol. 29 p. 1155-1168

McGurk, H., and MacDonald, J. (1976). “Hearing lips and seeing voices.” Nature, 264, 746-748.

Munhall, K G, Kroos, C., Jozan, G., and Vatikiotis-Bateson, E. (2004). "Spatial frequency requirements for audiovisual speech perception". *Perception & psychophysics (0031-5117)*, 66 (4), p. 574.

Shannon, R.V., Zeng, F., and Wygonski, J., (1998). “Speech recognition with altered spectral distribution of envelope cues.” Journal of the Acoustical Society of America 104, 2467-2476.

List of Figures/Tables

Percent Audiovisual Integration for Talkers Across Hearing Condition
Table 1

	Sloping	CI	Total
Ann	25.5	18.1	21.8
Christina	27.7	16.7	22.2
Bob	13.3	6.6	9.95
Joe	9.6	7.1	8.45

Audiogram of Sloping Hearing Loss (55dB HL at 1000 Hz) Condition in the HeLPS Program
Figure 1

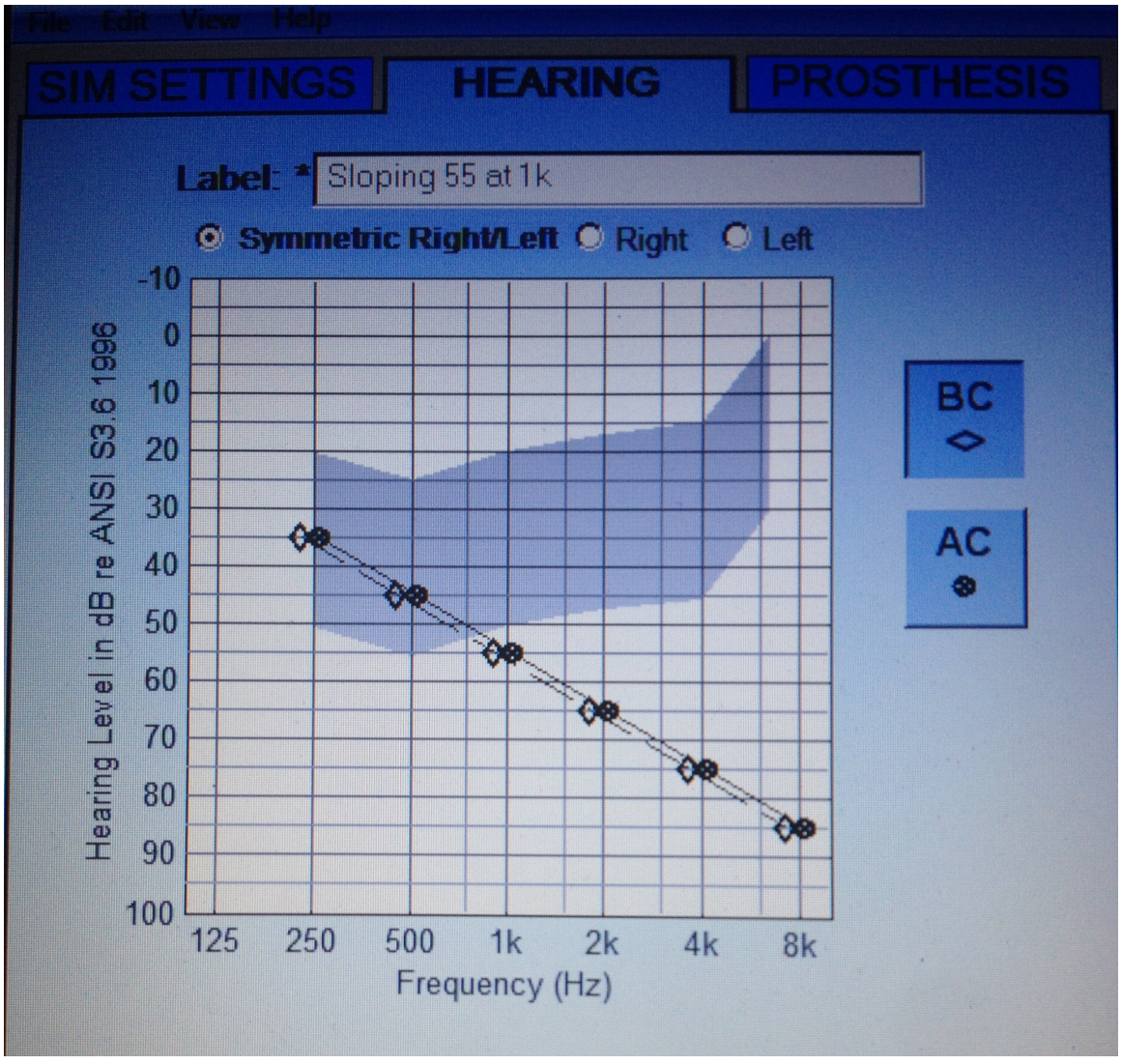

Figure 2. Percent key words correct for auditory, visual, and audio-visual modalities for Sloping and CI hearing conditions

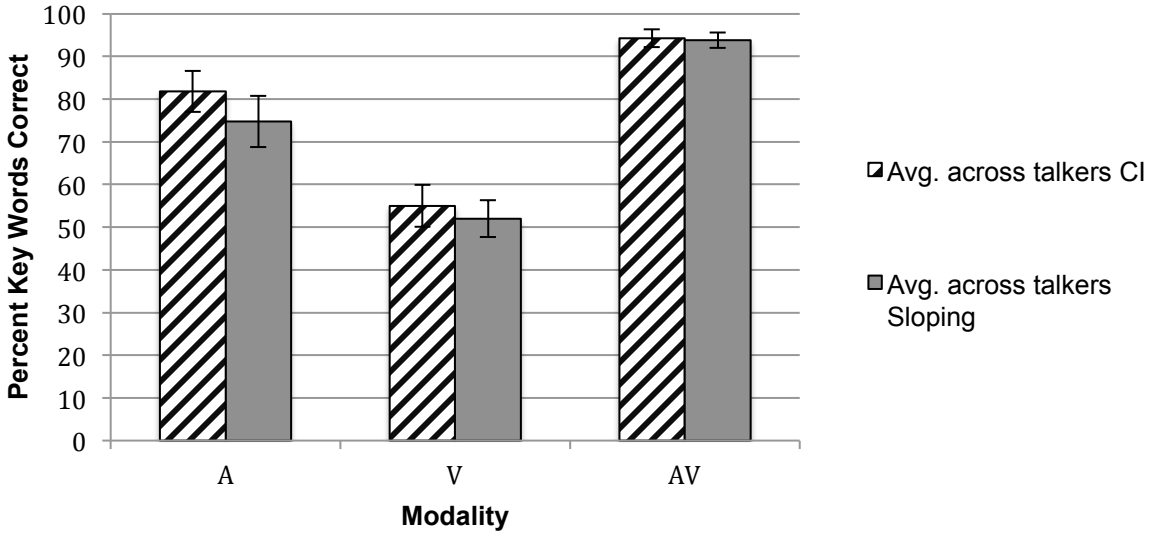

Figure 3. Percent key words correct for each talker in the Sloping hearing condition

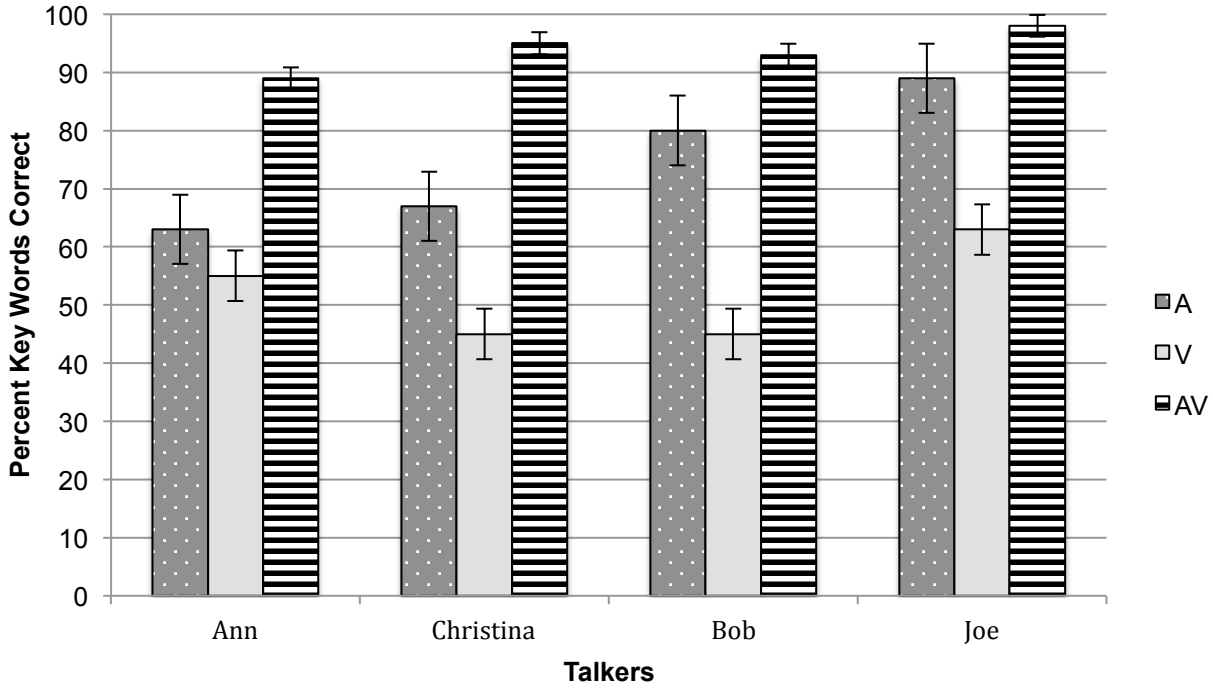

Figure 4. Percent key words correct for each talker for CI hearing condition

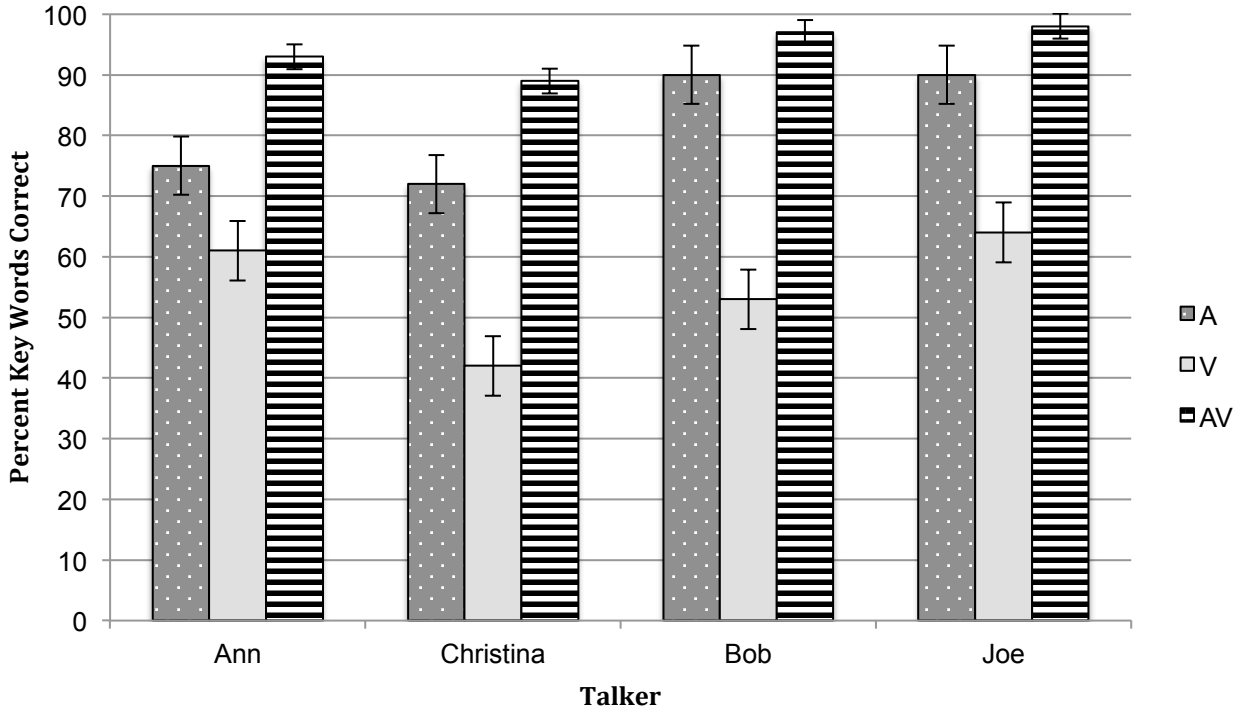

Figure 5. Average percent key words correct for each talker across modality and condition